IOP Conf. Series: Materials Science and Engineering **434** (2018) 012042 doi:10.1088/1757-899X/434/1/012042

The concept of sequential pattern mining for text

D S Maylawati[1]*, H Aulawi[2] and M A Ramdhani[3]

[1] Departement of Informatics, Sekolah Tinggi Teknologi Garut, Jalan Mayor Syamsu No 1 Tarogong Kidul Kabupaten Garut 44151, Indonesia
[2] Industrial Engineering, Sekolah Tinggi Teknologi Garut, Jalan Mayor Syamsu No 1 Tarogong Kidul Kabupaten Garut 44151, Indonesia
[3] Departement of Informatics, UIN Sunan Gunung Djati Bandung, Jalan A H Nasution No 105 Bandung 40614, Indonesia

*dsaadillah@sttgarut.ac.id

Abstract. Sequential pattern mining is one of popular data mining technique with sequential pattern as representation of data. However, most of sequential pattern mining research was conducted for structured data. In this paper, we did literature review of the sequential pattern mining algorithm that suitable for unstructured data such as text data. We reviewed several sequential pattern mining algorithm that had already used in text mining research, among others GSP, Spade, PrefixSpan, Spam, Lapin, SM-Spam, CM-Spade, BIDE, and another various algorithm based on sequential pattern mining problem such as concise representation and how to extract more rich pattern. The result showed that that from year to year research on text data using sequential pattern mining had increased. Although, not many algorithm were developed and also still rarely new algorithms were implemented in text data.

1. Introduction

Text data mining or text mining is a technique of gaining new knowledge from the text, automatically or semi-automatically, in which knowledge extracted from the text is useful and comes from a great number of text [1]–[3]. The data managed in the text mining technique is unstructurized text data. The difference between data mining and text mining is the extraction feature or the patterns coming from different forms of data. On the data mining, the extraction feature comes from a structured data, as for the text mining, it comes from a semi-structured data, and it cannot be considered as not structured at all or structured, although mostly come from the unstructured data [4], [5]. For that reason, the text data needs to be represented to be structured representation. In text mining, there is a pre-processing stage which prepares text data to become a structured representation [6], [7].

One of the structured representations of text is multiple of words. Multiple of words collect words on a text document by paying attention to the relations between words so that with the representation of multiple words, the semantic meaning on a text document can be maintained well, because it can understand the relations between words/phrases, even clauses and sentences [8], [9]. Sequential pattern mining (SPM), aside as one of the data mining techniques, can also be used as an algorithm to establish a structured text representation in the form of multiple of words. SPM algorithm results in sequential patterns between items. To this day, the SPM algorithm have developed well, such as GSP, Spade, PrefixSpan, Spam, Lapin, SM-Spam, CM-Spade, BIDE, and other SPM algorithms. The development of SPM algorithm is based on the needs of database that change and the efficiency of resulted pattern of

representation [10]. Basically, the SPM algorithm does the mining towards structured data. But, there also possibilities in which those algorithms are used in mining process that uses unstructured data, such as text data. In this research, we did a study on SPM algorithms literatures and did a simple survey about the use of those algorithms on text data.

2. Sequential pattern mining

Sequential pattern is a pattern established from a recurring transaction sequentially [4]. The technique of finding sequential pattern or sequential pattern mining is a part of data mining techniques that extracts the patterns in the form of sequential data, known as a set of features that stores the secrets information [10], [3], [11]. Sequential pattern emerges from the idea of transaction done in the supermarket which is usually done by the customers, there will be things that are usually bought simultaneously, there are also goods that are squentially bought after the others, so a pattern occurs [12]. But, although the sequence of the emergence is important, the emergence of an item on a sequential pattern doesn't have to be contiously, so there can be another item between one item and the others in a sequential pattern.

There are several terms in the sequential pattern, such as:

- Item set which is a a set of items that are not empty, for exampe an item set which is connoted with an i, in which $i = (i_j, i_{j+1}, i_{j+2}, \ldots, i_n)$ and i_j is an item.
- Then a sequence, i.e. a sequential list of several item sets, if the sequence is connoted with s, then $s = \{s_j, s_{j+1}, s_{j+2}, \ldots, s_n\}$ in which s_j is an item set. A sequence $A = \{a_1, a_2, a_3, \ldots, a_n\}$ is considered subsequence from sequence $B = \{b_1, b_2, b_3, \ldots, b_n\}$ and B is a supersequence from A, if the integers $i_1 < i_2 < i_3 < \ldots < i_n$ and $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i2}$, $a_3 \subseteq b_{i3}$, …, $a_n \subseteq b_{in}$. For example, a sequence {(3) (4, 5) (8)} is subsequence from a sequence {(7) (3, 8) (9) (4, 5, 6) (8)}, because (3) ⊆ (3, 8), (4, 5) ⊆ (4, 5, 6), and (8) ⊆ (8), but a sequence {(3) (5)} is not a subsequence from sequence {(3, 5)} and vice versa. This is because on sequence {(3) (5)} items emerge in order of one item set with another item set, meanwhile, on sequence {(3, 5)} items emerge simultaneously in an item set. In a set of sequence, a sequence is considered as maximum if a sequence is not a subsequence of another sequence.
- Furthermore, a support which is a value of a sequence that has to be met to indicate that number of transaction that consists of the sequence. A sequence is considered attractive if it meets the minimum support of the whole transaction. On the case of supermarket transaction, the support of this sequence is defined as a value of the total of customer's transaction which supports a sequence resulted

On the case of supermarket transaction, each transaction is considered as one item set and sorted based on the time of transactions, while all transactions of each customer is considered a sequence. The steps of finding the sequential pattern starts from table 1 and table 2, with a minimum support of 25% and applies the max function [12]. Table 1 is a transaction database table which has been sorted based on the transaction of each customer so it is known as sequence database. The transaction from table 1 is transformed to be a transformed database, as seen in table 2. Furthermore, the sequential pattern resulting from a transformed database on table 2 results in sequential pattern <(30) (90)> dan <(30) (40, 70)> with the emergence frequency of minimum twice of the whole transaction.

Table 1. The example of sequence database of transaction.

Customer ID	Transaction Time	Items Bought
1	June 25 '93	30
1	June 30 '93	90
2	June 10 '93	10, 20
2	June 15 '93	30
2	June 20 '93	40, 60, 70
3	June 25 '93	30, 50, 70
4	June 25 '93	30

Tabel 1. Cont.

4	June 30 '93	40, 70
4	July 25 '93	90
5	June 12 '93	90

Table 2. The example of sequence database of transaction.

Customer ID	Original Customer Sequence	Transformed Customer Sequence
1	<(30) (90)>	<{{(30)} {(90)}}>
2	<(10, 20) (30) (40, 60, 70)>	<{{(30)} {(40), (70), (40, 70)}}>
3	<(30, 50, 70)>	<{(30), (70)}>
4	<(30) (40, 70) (90)>	<{{(30)} {(40), (70), (40, 70)} {(90)}}>
5	<(90)>	<{(90)}>

2.1. Basic SPM algorithm and its variants

One of the SPM algorithms that is used the most is PrefixSpan, because the PrefixSpan reduces the database projection so that the performance is better than GSP, FreeSpan, and SPADE [13]. PrefixSpan is an SPM algorithm which adapts the divide and conquer principle and the pattern growth so that the mining process is more efficient, especially for the larger size of database [14]. Another SPM algorithm, such as GSP, which is adapted from Apriori algorithm in creating the sequential pattern [15]; SPADE algorithm which overcomes the problem of scanning database that happens repeatedly so that the mining process can become more efficient, although it still adapts the Apriori algorithm [16]; SPAM algorithm which integrates the depth first search with the depth first transversal techniques to make the sequential pattern which is too long to be efficient [17]; the FAST algorithm which represents the database with an indext sparse id-list to speed up the support calculation of the squential pattern [18]; CM-SPADE and CM-SPAN algorithms which uses the vertical representation to create a sequential pattern in which the process of generating candidate is accompanied by a test approach so it also rise the candidates which are not frequent to be eliminated [11]; also LAPIN algorithm which comes from the last position of an item so that it becomes the reference of whether the sequence resulted is frequent or not [19].

2.2. SPM Algorithm and its variants based on representation problem

There are three approaches to represent the frequent item set by selecting the features so that the frequent item set resulted can be more efficient. Those three approaches are maximum item set, close item set, and generator item set (key item set) approaches. A sequence s is said to be maximum if there is no more sequence s' which is a subsequence from a sequence s [8], [12], [20], [21]. For example, sequence s has items (a, b, c, d, e) and sequence s' has items (b, d, e), and both are frequent in the collection of documents. So that sequence s' is a subsequence from sequence s, so that sequence s is said to be maximum and sequence s' will be eliminated. Meanwhile, the close approach selects formed features to be more efficient. A sequence s is said to be close if there is no more sequence s' which is a subsequence from sequence s, in which sequence s and sequence s' have the same frequency [22], [23]. For example, sequence s has s (a, b, c, d , e) with the frequency of 3 and sequence s' (b, d, e) with the frequency of 3, then the sequence s is close and sequence s' will be eliminated. But, if sequence s' has a different frequency with sequence s which is its supersequence, then sequence s' will not be eliminated and will be included in close item sets. Last of all, the item set generator approach is the opposite of the close approach, if there is no more sequence s which is a supersequence from the subsequence s', in which sequence s and sequence s' have the same frequencies.

The SPM algorithm which applies the close principle, such as ClaSP, has an efficient searching method which uses the space pruning and vertical database layout [24]; the algorithm CloSpan which creates a sequential pattern from the large database [25]; the CloSpan algorithm and CM-ClaSpan which combines ClaSP and CloSpan [11]; CloFAST algorithm that combines sparse id-list and vertical id-list so that the calculation of the value of the support sequential pattern resulted is faster [26]; and the BIDE

algorithm which creates the sequential pattern without the process of candidate maintenance to result in frequent closed sequence [27]. The VMSP and Max SP algorithms are SPM algorithms that create the maximum sequential pattern. In which, VMPS is vertical mining of maximum sequential pattern which uses vertical mining the first time [28], as for MaxSP, it is an algorithm which uses maximum sequential pattern without doing the process of candidate maintenance [29]. Furthermore, the SPM algorithms which result in the sequential generator pattern are VGEN [30], FEAT [31], and FSGP [32].

2.3. Another variants of SPM Algorithm

Other SPM algorithms among others GoKrimp and SeqKrimp for compressing sequential patterns [33]; SeqDIM for frequent multidimensional sequential patterns derived from multidimensional sequence databases [34]; and the Songram algorithm for frequent closed multidimensional sequential patterns derived from multidimensional sequence databases [35]. There is also a Hirate-Yamana algorithm that performs SPM with min / max time interval between events and min / max time for
the length of the sequence [36]; and other Fournier-Viger Algorithms that combine SPM with dimensional pattern mining, time intervals, automatic clustering of valued actions, and closed sequence mining [37].

3. Sequential pattern mining for text

The representation of text with the result of SPM is known as frequent word sequence (FWS) [8], [9], [21]. FWS has a structure in which it sees documents or the collection of text datas as a set of frequent word sequence. The FWS structure can be illustrated by $\{(w_1,w_2), (w_3,w_4), \ldots\}$ in which (w_1,w_2) is FWS_i, (w_3,w_4) is FWS_{i+1}, etc. The order of the emergence of a set of FWS corresponds with the order of the emergence in the documents or a collection of text datas, as well as the emergence of elements on FWS is always in order with the emergence on a document or a collection of text datas. This means that in a collection of documents, FWS_i occurs frequently and followed by FWS_{i+1} and so on. As well as the elements or items in FWS_i, w_1 will always be followed by w_2, if w_2 emerges before w_1 it will be considered as a different FWS, as well as the emergence of elements in FWS_{i+1} and so on. Even the FWS has been developed not only by paying attention to the order of the words' emergences, but also by paying attention to the order of the words' emergences on the tex datas, known as a set of FWS [9].

Table 3. The example of document collection (presented in Indonesian Slang language).

No.	Content of document
1	*PKL di sekitar kawasan Gasibu sudah mulai ditertibkan. Para PKL masih boleh berjualan di area monumen sampai dengan di depan Telkom.*
2	*PKL di Gasibu telah ditertibkan untuk mengurangi kemacetan di sekitar Gasibu pada hari Minggu. Para PKL dibolehkan untuk berjualan di area Monumen.*
3	*Warga Bandung, mohon laporkan jika mendapati atau melihat PKL yang kembali berjualan lagi di kawasan yang sudah ditertibkan. Juga mohon laporkan jalan mana saja yang baru diaspal namun sudah rusak kembali.*

For example, from the example of the collection of documents in table 3, the representation of FWS which resulted on the minumum value of support of 50%, such as: {(pkl)}, {(pkl, tertib)}, {(pkl, jual)}, {(pkl, kawasan)}, {(pkl, gasibu, jual area)}, {(pkl, gasibu, tertib, jual)}, {(jual, area, monumen)}, {(area)}, {(gasibu)}, {(gasibu, jual)}, {(tertib, area, monumen)}, and {(gasibu, jual, area, monumen)}.

FWS {(pkl, tertib)} is different with {(tertib, pkl)} because those two FWS emerge in a different order on the collection of documents or text data.

4. Results and discussion

From a whole lot of SPM algorithm that develops continuously, we did a simple survey for each algorithm and saw the trend of SPM algorithm using for text data. We obtained the datas from Mendeley and Google Scholar, since those two indexing were representative and complete enough and represented the the researches published from various sources. Table 4 shows that from 23 SPM algorithm, PrefixSpan algorithm is the most used algorithm for a research with text datas, and there are 4 SPM algorithms which are implemented in a research using text datas, such as CM-SPAM, SPADE, BIDE, and GoKrimp algorithms. Meanwhile, the other 18 SPM algorithms have yet to be discovered in the research using the text datas. It shows that SPM algorithm has been used to search for sequential pattern from the unstructured datas, such as text, either on the text mining researh, information retrieval, or natural language processing. But, there are still a lot of SPM algorithms that have not been implemented on the research using text data.

Table 4. FIM algoithm for research with text data.

Algorithm	How many used for research with text data		
	0	**> 0 & < 5**	**≥ 5**
CM-SPADE	√		
CM-SPAM		√	
FAST	√		
GSP	√		
LAPIN	√		
PrefixSpan			√
SPADE		√	
SPAM	√		
ClaSP	√		
CM-ClaSP	√		
CloFAST	√		
CloSPAN	√		
BIDE		√	
VMSP	√		
MaxSP	√		
VGEN	√		
FEAT	√		
FSGP	√		
GoKrimp		√	
SeqKrimp	√		
SeqDIM	√		
Songram	√		
Hirate-Yamana	√		

5. Conclusion

SPM is one of the data mining techniques which looks for the sequential pattern from the transaction database. Basically, SPM is used to do a mining for structurised datas. But, SPM can also be used on the unstructurised datas, such as text, which results in FWS as a structured representation of text. From the whole lot of SPM algorithm developed, there is only 1 out of 23 algorithms or 4.34% that is used in the research using text datas, 17.39% (4 out of 23) of SPM algorithms are used on the research using text datas, although there are just a few. Meanwhile, 78.27% (18 out of 23) others have not been

implemented on a text data. This becomes the opportunity for an upcoming research in implementing and researching on the SPM algorithms for text datas, both on text mining techniques, information retrieval, and natural language processing.

Acknowledgement
We would like to thank *Sekolah Tinggi Teknologi Garut* for the supports; both moral and material supports so that this article could be published.

References

[1] C J Torre, M J Martin Bautista, D Sanchez and I Blanco 2008 Text Knowledge Mining: And Approach To Text Mining *ESTYLF08* 17–19
[2] V Gupta and G S Lehal 2010 A Survey of Text Summarization Extractive techniques in *Journal of Emerging Technologies in Web Intelligence* **2**(3) pp. 258–268
[3] H Jiawei, M. Kamber, J Han, M Kamber and J Pei 2012 Data Mining: Concepts and Techniques Elsevier
[4] H Jiawei, M Kamber, J Han, M Kamber and J Pei 2006 Data Mining: Concepts and Techniques Elsevier
[5] S M Weiss, N Indurkhya, T Zhang and F J Damerau 2010 Information Retrieval and Text Mining (Springer Berlin Heidelb) Fundamentals of Predictive Text Mining, pp. 75–90
[6] H Mahgoub, D Rösner, N Ismail and F Torkey 2008 A Text Mining Technique Using Association Rules Extraction *Int. J. Comput. Intell.* **4**(1) pp. 21–28
[7] D Sa'Adillah Maylawati and G A Putri Saptawati 2017 Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang in *Journal of Physics: Conference Series* **801(**1)
[8] A Doucet and H Ahonen Myka 2004 Non-contiguous word sequences for information retrieval *MWE '04 Proc. Work Multiword Expressions* **26** pp. 88–95
[9] D S A Maylawati 2015 Pembangunan library pre-processing untuk text mining dengan representasi himpunan frequent word itemset (HFWI) *Studi Kasus: Bahasa Gaul Indonesia* (Bandung: ITB)
[10] P Fournier Viger, J Chun, Wei Lin, R U Kiran, Y S Koh and R Thomas A Survey of Sequential Pattern Mining *Ubiquitous Int.* **1**(1) pp. 54–77
[11] P Fournier Viger, A Gomariz, M Campos and R Thomas 2014 Fast vertical mining of sequential patterns using co-occurrence information in *Pacific-Asia Conference on Knowledge Discovery and Data Mining* pp. 40-52 (Cham: Springer)
[12] R Agrawal and R. Srikant 1994 Fast Algorithms for Mining Association Rules in Large Databases *J. Comput. Sci. Technol.* **15(**6) pp. 487–499
[13] D S A Maylawati, M A Ramdhani, A Rahman and W Darmalaksana Incremental technique with set of frequent word item sets for mining large Indonesian text data in *2017 5th International Conference on Cyber and IT Service Management CITSM 2017*
[14] Pei J, Han J, Mortazavi Asl B, Wang J, Pinto H, Chen Q and Hsu M C 2004 Mining sequential patterns by pattern-growth: The prefixspan approach *IEEE Transactions on Knowledge & Data Engineering* (11) 1424-1440.
[15] R Srikant and E Agrawal 1996 Mining Sequential Patterns: Generalization and Performance Improvements *5th Int. Conf. Extending Database Technol (EDBT '96)* pp. 3–17
[16] M J Zaki 2001 SPADE: An efficient algorithm for mining frequent sequences *Mach. Learn.* **42(**1–2) pp. 31–60
[17] J Ayres, J Flannick, J Gehrke and T Yiu 2002 Sequential PAttern mining using a bitmap representation in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02* p. 429
[18] E Salvemini, F Fumarola, D Malerba and J Han 2011 Fast sequence mining based on sparse id-lists in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial*

Intelligence and Lecture Notes in Bioinformatics) 2011 6804 pp. 316–325
[19] Z Yang, Y Wang and M Kitsuregawa 2007 LAPIN: effective sequential pattern mining algorithms by last position induction for dense databases *12th Int. Conf. Database Syst. Adv. Appl. DASFAA 2007* **1** pp. 1020–1023
[20] R Agrawal, H Mannila, R Srikant, H Toivonen and a I Verkamo 1996 Fast discovery of association rules *Advances in knowledge discovery and data mining* **12** pp. 307–328
[21] H Ahonen Myka 2002 Discovery of Frequent Word Sequences in Text *Proc. ESF Explor. Work. Pattern Detect.Discov.* pp. 180–189
[22] J Wang, J Han and J Pei 2003 Closet+: Searching for the best strategies for mining frequent closed itemsets *Proc. ninth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.* pp. 236–245
[23] T Uno, T Asai, Y Uchida and H Arimura 2003 LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets *Fimi* **90**
[24] Gomariz A, Campos M, Marin R and Goethals B 2013 Clasp: An efficient algorithm for mining frequent closed sequences In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* pp. 50-61 (Berlin Heidelberg: Springer)
[25] X Yan, J Han and R Afshar 2003 CloSpan: Mining Closed Sequential Patterns in Large Datasets *Conf. Proc. Third SIAM Int. Conf. Data Mining* (USA: San Fr. CA)
[26] F Fumarola, P F Lanotte, M Ceci and D Malerba 2016 CloFAST: closed sequential pattern mining using sparse and vertical id-lists *Knowl. Inf. Syst.* **48**(2) pp. 429–463
[27] J Wang and J Han 2004 BIDE: Efficient Mining of Frequent Closed Sequences in *Data Engineering 2004 Proceedings 20th International Conference*
[28] Fournier Viger P, Wu C W, Gomariz A and Tseng V S 2014 VMSP: Efficient vertical mining of maximal sequential patterns In *Canadian Conference on Artificial Intelligence* (Cham: Springer) pp. 83-94
[29] Fournier Viger P, Wu C W and Tseng V S 2013 Mining maximal sequential patterns without candidate maintenance In *International Conference on Advanced Data Mining and Applications* (Berlin Heidelberg: Springer) pp. 169-180
[30] Fournier Viger P, Gomariz A, Šebek M and Hlosta M 2014 VGEN: fast vertical mining of sequential generator patterns In *International Conference on Data Warehousing and Knowledge Discovery* (Cham: Springer) pp. 476-488
[31] C Gao, J Wang, Y He and L Zhou 2008 Efficient mining of frequent sequence generators *Proceeding 17th Int. Conf. World Wide Web - WWW '08* p. 1051
[32] S Yi, T Zhao, Y Zhang, S Ma and Z Che 2011 An effective algorithm for mining sequential generators in *Procedia Engineering* **15** pp. 3653–3657
[33] H T Lam, F Mörchen, D Fradkin and T Calders 2014 Mining Compressing Sequential Patterns *Stat. Anal. Data Min.***7**(1) pp. 34–52
[34] H Pinto, J Han, J Pei, K Wang, Q Chen and U Daya 2001 Multi-dimensional sequential pattern mining in *Proceedings of the tenth international conference on Information and knowledge management CIKM'01* p. 81.
[35] P Songram, V Boonjing and S Intakosum 2006 Closed multidimensional sequential pattern mining in *Proceedings Third International Conference onInformation Technology: New Generations ITNG 2006* pp. 512–517
[36] Y Hirate and H Yamana 2006 Generalized sequential pattern mining with item intervals *J. Comput.***1**(3) pp. 51–60
[37] Fournier Viger P, Nkambou R and Nguifo E M 2008 A knowledge discovery framework for learning task models from user interactions in intelligent tutoring systems In *Mexican International Conference on Artificial Intelligence* (Berlin Heidelberg: Springer) pp. 765-778